The dynamic latent block model for the co-clustering of evolving binary matrices

Giulia Marchello[1,2], Marco Corneli[2,3] & Charles Bouveyron[2]

[1] *Università degli Studi di Perugia*
[2] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai team*
[3] *Université Côte d'Azur, Maison de la Simulation et des Interactions (MSI)*
Email: giulia.marchello@inria.fr

Résumé. Nous considérons le problème du co-clustering des matrices binaires qui peuvent évoluer dans le temps et nous introduisons un modèle génératif pour le gérer. Le modèle proposé, appelé dynamic latent block model, étend le modèle des blocs latents binaire classique au cas dynamique. La modélisation de la dynamique en temps continu repose sur un processus de Poisson non homogène, avec une partition latente des intervalles de temps. Nous proposons d'utiliser l'algorithme SEM-Gibbs pour l'inférence du modèle.

Mots-clés. Co-clustering, matrices binaires dynamiques, modèle à blocs latents, algorithme SEM-Gibbs

Abstract. We consider the problem of co-clustering binary matrices that may evolved along the time and we introduce a generative model to handle it. The proposed model, named dynamic latent block model, extend the classical binary latent block model to the dynamic case. The modeling of the dynamic in a continuous time relies on a non-homogeneous Poisson process, with a latent partition of time intervals. We proposed to use the SEM-Gibbs algorithm for model inference.

Keywords. Co-clustering, dynamic binary matrices, latent block model, SEM-Gibbs algorithm

1 Introduction

In many applications, it is now frequent to have to summarize large binary matrices that may evolve along the time. For instance, e-commerce systems record in continuous time all purchases of products made by customers. It is of interest for those companies to cluster both customers and products to better understand the purchasing behaviors. The simultaneous clustering of rows and columns of a matrices is known as a co-clustering problem.

The Dynamic Latent Block Model introduced in the present work is a new and original way to consider, in a dynamic framework, the Latent Block Model [see e.g. [2, 7]] which is a popular generative model for co-clustering. We assume that the number of individuals

(rows) $i = 1, ..., n$ and the number of objects (columns) $j = 1, ..., p$ is fixed during the whole time period taken into account $[0, T]$. Moreover, we consider that the interaction occurring between the individual i and the object j at time $u \in [0, T]$, is represented by the triplet (i, j, u) with $u \leq T$, and it is denoted by $x_{ij}(u)$. Thus, we assume that the number of interactions between individuals and objects follows a non-homogenous Poisson process (NHPP) where the intensity function $\lambda(t)$ only depends on the clusters they belong to [4].

In order to model the continuous time we consider the approach that has been used in Corneli et al. (2018) [3]. In that paper the continuous time is handled by a time partition over $[0, T]$ where the interactions are aggregated on the time intervals of such partition obtaining a sequence of static matrices that allows us to identify the time clusters. Therefore, the basic idea is to partition the time in small windows and to cluster every time period, in this way it is possible to assign every time interval to a time cluster. Intuitively, one time cluster can occur more than once in the temporal line when its peculiar features are repeated after some time. This behaviour, for instance, is typical when the phenomenon of seasonality is observed.

2 The dynamic latent block model

We consider a random matrix $\mathbf{X}(t)$ where the entry $X_{ij}(t)$, with $i = 1, ..., n$, $j = 1, ..., p$ and $t \in [0, T]$, counts the number of interactions between i and j up to time t.

The latent structure of rows and columns of the matrix $\mathbf{x}(t)$ is identified by:

- $\mathbf{z} = (z_{ik}; i = 1, ..., n; k = 1, ..., K)$: represents the clustering of rows into K groups: $\mathcal{A}_1, ..., \mathcal{A}_K$. Row i belongs to cluster $\mathcal{A}_k$ iff $z_{ik} = 1$ and $\mathbf{z_i} = (z_{ik})_k \in (0, 1)^K$ is the group idicator of row i;
- $\mathbf{w} = (w_{j\ell}; j = 1, ..., p; \ell = 1, ..., L)$: represents the clustering of columns into L groups: $\mathcal{B}_1, ..., \mathcal{B}_L$. Coloumn j belongs to cluster B_ℓ iff $w_{j\ell} = 1$ and $\mathbf{w_j} = (w_{j\ell})_\ell \in (0, 1)^L$ is the group idicator of coloumn j;

Moreover, $\mathbf{z}$ and $\mathbf{w}$ are indipendent and they are distributed as follows:

$$p(\mathbf{z}|\gamma) = \prod_{k=1}^{K} \gamma_k^{|\mathcal{A}_k|} \tag{1}$$

where:
$\gamma_k = \mathbb{P}\{z_i = k\}$; $\sum_{k=1}^{K} \gamma_k = 1$ and $|\mathcal{A}_k|$ represents the number of rows in the cluster $\mathcal{A}_k$.

$$p(\mathbf{w}|\rho) = \prod_{\ell=1}^{L} \rho_\ell^{|\mathcal{B}_\ell|} \tag{2}$$

where:

$\rho_\ell = \mathbb{P}\{w_j = \ell\}$; $\sum_{\ell=1}^{L} \rho_\ell = 1$ and $|\mathcal{B}_\ell|$ represent the number of coloums in the cluster $\mathcal{B}_\ell$.

As mentioned previously, a non-homogeneous Poisson process (NHPP) is used to count the interactions between the row i and the column j up to time $t \in [0, T]$, denoted by $X_{ij}(t)$:

$$X_{ij}(t)|z_{ik}w_{j\ell} = 1 \sim \mathcal{P}\left(\int_0^t \lambda_{k\ell}(u)du\right) \tag{3}$$

where $\lambda_{k\ell}(t)$ represents the intensity function that only depends on the considered row cluster k and on the column cluster ℓ. Moreover, $\lambda_{k\ell}(t)$ has to be positive and integrable on the time interval $[0, T]$.

As in Corneli (2016) [4], we discretize the continous time interval [0,T] in U subintervals, where $I_u = [t_{u-1} - t_u[$. , with:

$$0 = t_0 < t_1 < \cdots < t_U = T.$$

The number of interactions between i and j on the considered time partition I_u is summarized by X_{iju} and is defined as:

$$X_{iju} := X_{ij}(t_u) - X_{ij}(t_{u-1}), \forall (i, j, u).$$

We introduce a tensor $X = \{X_{iju}\}_{iju}$ with dimensionality $N \times P \times U$. As previously described in a more theoretical way, each time interval $I_1, ..., I_U$ is assigned to a hidden time cluster $\mathcal{C}_1, .., \mathcal{C}_c$. To model the membership to time clusters, a new latent variable $\mathbf{s}$ has to be introduced, in particular $\mathbf{s}_u = c$ if and only if the time interval I_u belongs to the time cluster $\mathcal{C}_c$. Furthermore, we assume that $\mathbf{s}$ follows a multinomial distribution:

$$p(\mathbf{s}|\delta) = \prod_{c=1}^{C} \delta_c^{|\mathcal{C}_c|}, \tag{4}$$

where:

$\delta_c = \mathbb{P}\{s_u = c\}$; $\sum_{c=1}^{C} \delta_c = 1$ and $|\mathcal{C}_c|$ represents the number of time intervals in the cluster $\mathcal{C}_c$. The originality of this work lies in the fact that a dependence is assumed between the two variables: X_{iju} and S_u.

Once these additional assumpions have been made, we can rewrite Eq; (3) considering that the intensity functions are stepwise constant on each time cluster $\mathcal{C}_c$. Hence, now we are considering a non- homogeneus Poisson process (NHPP) with parmeter $\lambda_{k\ell c}$ and Δ_u :

$$X_{iju}|z_{ik}w_{j\ell}S_{uc} = 1 \sim \mathcal{P}(\lambda_{k\ell c}\Delta_u) \tag{5}$$

where Δ_u indicates the length of the interval I_u that is usually constant, $\Delta_u = \Delta$. Moreover, as pointed out in Corneli et al. (2018) [3], one can set $\Delta_u = 1$ without loss of generality, indeed if, for instance, we have a dataset divided in week, the time interval Δ_u could be set to one week. It is finally possible to introduce a tensor $\Lambda = \{\lambda_{k\ell c}\}_{k\ell c}$ of dimension $K \times L \times C$.

3 The likelihood

Since the increments of a Poisson process are independent it holds that:

$$p(X_{iju}|z_{ik}w_{jl}s_{uc} = 1, \lambda_{k\ell}) = \prod_{u=1}^{U} \left(\frac{(\lambda_{klc})^{X_{iju}}}{X_{iju}!} \exp\left(-\lambda_{klc}\right) \right) \tag{6}$$

Therefore, we can introduce the $K \times L \times C$ tensor of order 3, $\boldsymbol{\lambda}$, identified by the triplet (i, j, u) and whose elements are denoted as $\pi_{k\ell c}$. At this point it is possible to write the complete data likelihood of the model, identified by the following equation:

$$p(X, z, w, s|\gamma, \rho, \delta, \lambda) = p(z|\gamma) \cdot p(w|\rho) \cdot p(s|\delta) \cdot p(X|z, w, s, \lambda) \tag{7}$$

Looking at the right hand side of the Eq. (7) we notice that $p(z|\gamma)$, $p(w|\rho)$ and $p(s|\delta)$ have been defined in the previous section respectively by the Eqs. (1), (2) and (4). While the joint distribution of $\mathbf{X}$, given $\boldsymbol{z}$, $\boldsymbol{w}$, and $\boldsymbol{s}$, can be easily obtained as a generalization of the Eq. (6):

$$p(X|z, w, s, \lambda) = \prod_{k=1}^{K}\prod_{\ell=1}^{L}\prod_{c=1}^{C} \left(\frac{(\lambda_{k\ell c})^{R_{k\ell c}}}{P_{k\ell c}} exp\left(-|\mathcal{A}_k||\mathcal{B}_\ell||\mathcal{C}_c|\lambda_{k\ell c}\right) \right) \tag{8}$$

where:

$$R_{k\ell c} = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{u=1}^{U} z_{ik}w_{j\ell}s_{uc}X_{iju}$$

$$P_{k\ell c} = \prod_{i=1}^{n}\prod_{j=1}^{p}\prod_{u=1}^{U} (z_{ik}w_{j\ell}s_{uc}X_{iju})!$$

Denoted as θ the set of the model parameters: $\theta = (\gamma, \rho, \delta, \lambda)$. The complete data log-likelihood can be defined as

$$\ell(\theta|z, w, S, X) = \sum_{z}\sum_{w}\sum_{c} \log p(X, z, w, S|\theta) \tag{9}$$

4 Model inference

As usual, we look for a way to maximize the log-likelihood in order to obtain the estimation of $\boldsymbol{\theta}$. In the co-clustering case, the EM algorithm is computationally infeasible [6], the idea is to use a stochastic version of it, known as SEM-Gibbs, proposed by Keribin C. (2010) [9] and exploited, for instance, by Bouveyron et al. (2017) [1] in the Functional Latent Block Model . Thanks to the Gibbs Sampling, in the SE step a partition for $\boldsymbol{z}$, $\boldsymbol{w}$ and $\boldsymbol{s}$ is generated without computing the joint distribution. The algorithm starts with initial values for the parameter set $\boldsymbol{\theta}^{(0)}$, the column clusters $\boldsymbol{w}^{(0)}$ and the time clusters $s^{(0)}$. Regarding the burn-in period, after a certain number of iterations of the algorithm, we can obtain the final parameters estimation by computing the mean of the sampled distribution. The optimal values for $\boldsymbol{z}$, $\boldsymbol{w}$ and $\boldsymbol{s}$ are estimated by the mode of their sample distributions.

5 Conclusion and further work

We introduced in this short paper a generative model for co-clustering dynamic binary matrices. We proposed to use the SEM-Gibbs algorithm for model inference. Regarding further work, we are currently working on an implementation of this algorithm and we plan to apply it on real-world data from Amazon recording fine food purchases over 8 years.

References

[1] Charles Bouveyron, Laurent Bozzi, Julien Jacques, and François-Xavier Jollois. The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4):897–915, 2018.

[2] Vincent Brault and Mahendra Mariadassou. Co-clustering through latent bloc model: A review. 2015.

[3] Marco Corneli, Charles Bouveyron, Pierre Latouche, and Fabrice Rossi. The dynamic stochastic topic block model for dynamic networks with textual edges. *Statistics and Computing*, 2018.

[4] Marco Corneli, Pierre Latouche, and Fabrice Rossi. Block modelling in dynamic networks with non-homogeneous poisson processes and exact icl. *Social Network Analysis and Mining*, 6(1):55, 2016.

[5] Gérard Govaert and Mohamed Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245, 2008.

[6] Gérard Govaert and Mohamed Nadif. *Co-clustering: models, algorithms and applications*. John Wiley & Sons, 2013.

[7] Christine Keribin, Vincent Brault, Gilles Celeux, Gérard Govaert, et al. Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, volume 2012, 2012.

[8] Christine Keribin, Gilles Celeux, and Valérie Robert. The Latent Block Model: a useful model for high dimensional data. In *ISI 2017 - 61st world statistics congress*, pages 1–6, Marrakech, Morocco, July 2017.

[9] Christine Keribin, Gérard Govaert, and Gilles Celeux. Estimation d'un modèle à blocs latents par l'algorithme sem. 2010.